© J.-B.A.K. <jean-baptiste.apoung@math.u-psud.fr>

Calcul scientifique au service d'une équation aux dérivées partielles

L'objectif de ce projet est de familiariser l'étudiant avec certaines facettes peut-être cachées du calcul scientifique. En l'occurrence, certains problèmes qui préoccupent le spécialiste du calcul scientifique et les choix de résolution. Au regard du niveau des étudiants concernés, un choix délibéré est fait de ne pas s'appesantir sur la facette analyse numérique du calcul scientifique, encore moins sur la facette modélisation de cette discipline.
Unités d'enseignement requises : Math 315 et Math 325.

Partie - 1 *Un problème modèle*

On considère le problème suivant :

$$\partial_t u - \varepsilon \partial^3_{xxt} u + \partial_x f(u) + \mu \partial^3_{xxx} u = 0 \quad \text{in} \quad]0,T[\times\Omega, \tag{1}$$

où $\varepsilon, \mu, a, b, T > 0$ sont des nombres réels donnés et $\Omega =]a,b[$ et $\partial_t, \partial^3_{xxt}$ et ∂^3_{xxx} désignent respectivement, les dérivées partielles d'ordre 1 en temps, d'ordre trois dont deux en espace et un en temps, et d'ordre trois en espace. On considère les conditions initiales et aux limites suivantes

$$u(0,x) = u_0(x) \quad \forall x \in [a,b], \tag{2}$$

$$\left\{ \begin{array}{rcll} u(t,a) & = & u(t,b) & \forall t \in [0,T], \\ u_x(t,a) & = & u_x(t,b) & \forall t \in [0,T], \\ u_{xx}(t,a) & = & u_{xx}(t,b) & \forall t \in [0,T]. \end{array} \right. \tag{3}$$

où le terme non linéaire a la forme suivante :

$$f(u) = \beta u + \vartheta \frac{u^2}{2}, \tag{4}$$

où β, ϑ sont des nombres positifs. Les paramètres dispersifs $\varepsilon \geq 0$ et $\mu \geq 0$ ne s'annulent pas simultanément. Lorsque $\varepsilon \neq 0$ et $\mu = 0$ l'équation est dite de Benjamin Bonas Mahoney (BBM) et lorsque $\varepsilon = 0$ and $\mu \neq 0$ l'équation est dite de Korteweg De Vries (KDV).

On s'assure qu'on s'attaque bien à un problème résoluble :

La littérature nous assure de l'existence, l'unicité et d'une régularité suffisante de la solution de cette équation. Nous n'allons pas nous y appesantir, et retiendrons simplement que l'intégration en espace sur des domaines bornés, d'une solution de cette équation est possible et qu'il en résultera une fonction continue et dérivable en temps. Il convient néanmoins de s'en assurer pour plus de crédibilité.

Q-1**-**1 **:** Rechercher et inclure dans votre rapport une ou deux références bibliographiques dans lesquelles sont évoquées l'existence et l'unicité de la solution à ce type d'équations aux dérivées partielles.

On identifie des caractéristiques particulières du problème, nécessaires au niveau numérique :

Cette équation possède des solutions particulières dites ondes solitaires d'expression :

$$u(t,x;C_s,D_s) = 3(C_s-1)sech^2\Big(\sqrt{\frac{C_s-1}{4(\mu+\varepsilon C_s)}}\,(x - C_s\,t - D_s)\Big), \tag{5}$$

où $sech(x) = 1/\cosh(x)$ *avec* $\cosh(x) = \frac{e^x+e^{-x}}{2}$*. Elles servent de base de validation car des méthodes numériques destinées à résoudre ces équations dans des domaines plus complexes, doivent en particulier fournir des approximations raisonnables de ces ondes solitaires.*

Dans le même ordre d'idées, notons que l'équation (1) possède certaines propriétés qu'il serait important, dans la limite du possible, de préserver à travers le schéma numérique.

Q-1**-**2 **:** Montrer que pour la solution u du problème (1) les quantités suivantes sont conservées au cours du temps :

$$I_1(t) = \int_a^b u(t,x)\,dx, \quad \text{et} \quad I_2(t) = \int_a^b \Big(u^2(t,x) + \varepsilon u_x^2(t,x)\Big)\,dx \tag{6}$$

Partie - 2 *Discrétisation spatiale*

L'objectif est de s'affranchir de la dimension spatiale de sorte à passer de l'équation aux dérivées partielles à un système d'équations différentielles ordinaires. C'est la **méthode des lignes***.*

On choisit une discrétisation simple et plus à même de préserver certaines propriétés de la solution exacte.

Q-2-1 **:** On introduit de nouvelles variables intermédiaires suivantes :

$$r = \partial_x u, \qquad q = \partial_x r. \tag{7}$$

auxquelles on associe encore des conditions aux limites périodiques.

Montrer que le problème (1)– (2) devient

$$\left\{ \begin{array}{rcl} \partial_t(u - \varepsilon q) + \partial_x\Big(f(u) + \mu q\Big) & = & 0, \\ q - \partial_x r & = & 0, \\ r - \partial_x u & = & 0, \\ u(0,\cdot) & = & u_0, \end{array} \right. \tag{8}$$

Q-2-2 **:** On introduit une subdivision de l'intervalle $[a, b]$, appelée **un maillage** et désignée par τ_h. Elle est définie par

$$a = x_1 < x_1 < \ldots < x_j < x_{j+1} < \ldots < x_{N+1} = b.$$

On pose alors $I_j = [x_j, x_{j+1}] \quad \forall j = 1, \ldots, N$ et on désigne par h_j la longueur de l'intervalle I_j. Par définition $\tau_h = \{I_j, j = 1, \ldots, N\}$, dans laquelle $h = \max_{1 \le j \le N} h_j$.

On convient ensuite de la manière dont la solution sera exprimée dans chaque intervalle I_j. Cela va conduire à différents types de méthodes. Pour le problème présent, nous faisons le choix d'utiliser une méthode de **volumes finis** pour approcher les solutions. Cette méthode se fonde sur la considération qu'une valeur approchée d'une fonction v sur l'intervalle I_j est donnée par sa valeur moyenne c'est-à-dire :

$$v_j = \frac{1}{h_j} \int_{x_j}^{x_{j+1}} v(x)\, dx. \tag{9}$$

Ce qui conduit à une approximation de v sur le maillage τ_h, donnée par

$$v_h(x) = \sum_{j=1}^{N} v_j \chi_{I_j}(x). \tag{10}$$

où $\chi_{I_j}(x)$ est la fonction caractéristique de l'intervalle (**maille**) I_j. C'est-à-dire : $\chi_{I_j}(x) = \left\{ \begin{array}{l} 1 \text{ si } x \in I_j \\ 0 \text{ sinon.} \end{array} \right.$

En intégrant l'équation (8) sur la maille I_j, montrer que l'on a :

$$\left\{ \begin{array}{rcl} \dfrac{d}{dt}\Big(\displaystyle\int_{I_j} (u - \varepsilon q)\, dx\Big) + \Big(f(u(t, x_{j+1})) + \mu q(t, x_{j+1})\Big) - \Big(f(u(t, x_j)) + \mu q(t, x_j)\Big) & = & 0, \\ \displaystyle\int_{I_j} q(t, x)\, dx - r(t, x_{j+1}) + r(t, x_j) & = & 0, \\ \displaystyle\int_{I_j} r(t, x)\, dx - u(t, x_{j+1}) + u(t, x_j) & = & 0, \\ \displaystyle\int_{I_j} u(0, x)\, dx & = & \displaystyle\int_{I_j} u_0(x)\, dx. \end{array} \right. \tag{11}$$

C'est le problème semi-discret **continu***. Il reste à en déduire un problème semi-discret* **approché***. Cela revient à faire apparaître dans les équations (11) les quantités u_j, r_j, q_j, approximations de u, r, q sur l'intervalle I_j.*

Dans le cadre de l'approximation par volumes finis, l'approximation des quantités volumiques se fait de manière naturelle mais par contre celle des quantités frontières est plus délicate. Plusieurs choix sont possibles, nous adoptons les suivants :

$$\left\{\begin{array}{rcl} \forall j=1,\quad \ldots,\quad N, & & \\ u_j(t) & = & \dfrac{1}{h_j}\displaystyle\int_{I_j} u(t,x)\,dx \\ q_j(t) & = & \dfrac{1}{h_j}\displaystyle\int_{I_j} q(t,x)\,dx \\ r_j(t) & = & \dfrac{1}{h_j}\displaystyle\int_{I_j} r(t,x)\,dx \\ q(t,x_j) & = & q_j(t) \\ r(t,x_j) & = & r_j(t) \\ u(t,x_j) & = & u_{j-1}(t) \\ f(u(t,x_j)) & = & \dfrac{f(u_{j-1}(t))+f(u_j(t))}{2} - \dfrac{\kappa\Big(u_{j-1}(t),u_j(t)\Big)}{2}\Big(u_j(t)-u_{j-1}(t)\Big). \end{array}\right. \tag{12}$$

où

$$\kappa(v,w) = \max_{\min(v,w)\le s\le\max(v,w)} |f'(s)|. \tag{13}$$

Les conditions aux limites périodiques doivent être prises en compte si nécessaire.

Q-2-3 : En portant ces expressions (12) dans (11), écrire le problème semi-discrétisé approché en espace.

Partie - 3 *Formation du système d'EDO*

On récrit le problème de sorte à tirer profit des outils à disposition (comme les solveurs d'ODE de Matlab)

Il est question de ramener le problème semi-discrétisé obtenu précédemment sous la forme d'un système d'équations différentielles ordinaires (sur la seule inconnue u si possible). C'est-à-dire sous la forme

$$\left\{\begin{array}{rcl} \dfrac{dU(t)}{dt} & = & F(t,U(t)), \quad t\in]0,T[, \\ U(0) & = & U^0. \end{array}\right. \tag{14}$$

où U^0 est à déterminer et $U(t)$ est le vecteur des $u_i(t), j=1,\ldots,N$ et $F(\cdot,\cdot)$ est une fonction de $\mathbb{R}\times\mathbb{R}^N$ dans $\mathbb{R}^N$.

On pose

$$\left\{\begin{array}{rcl} U(t) & = & [u_1(t),\ldots,u_N(t)]^T, \\ R(t) & = & [r_1(t),\ldots,r_N(t)]^T, \\ Q(t) & = & [q_1(t),\ldots,q_N(t)]^T, \end{array}\right. \tag{15}$$

et on désigne par M la matrice diagonale de taille N dont le i-ème terme diagonal est h_i : on l'appelle aussi **matrice de masse**.

Q-3-1 : Déterminer les matrices carrées L_r, L_q toutes deux de taille N telles que

$$R(t) = L_r U(t) \tag{16}$$

$$Q(t) = L_q R(t). \tag{17}$$

Q-3-2 : En déduire l'existence d'un opérateur $\mathcal{L}_{rq} : \mathbb{R}^N \mapsto \mathbb{R}^N$ et d'un vecteur $U^0 \in \mathbb{R}^N$ tels que le problème discret en espace obtenu précédemment se mette sous la forme

$$\left\{\begin{array}{rcl} (M-\varepsilon M L_q L_r)\dfrac{dU(t)}{dt} & = & L_{rq}(U(t)), \quad t\in]0,T[, \\ U(0) & = & U^0. \end{array}\right. \tag{18}$$

Q-3-3 : En déduire l'expression de $F(\cdot,\cdot)$ pour que le problème semi-discrétisé en espace soit de la forme (14).

Partie - 4 *Vers le calcul scientifique à proprement parler : le calcul efficace de $F(t,V)$*

Il est question pour le spécialiste du calcul scientifique de s'assurer que l'évaluation de F n'épuisera pas les ressources (espace mémoire et temps de calcul) mises à sa disposition.

Le problème ici consiste à évaluer efficacement l'expression $F(t,V) = (M - \varepsilon M L_q L_r)^{-1} L_{rq}(V)$.

On pose alors

$$\mathcal{A} = M - \varepsilon M L_q L_r. \tag{19}$$

On s'évertue à réduire au maximum le coût en espace mémoire.

Q-4-1 : Est-il raisonnable de calculer explicitement $\mathcal{A}^{-1}$?

Q-4-2 : En analysant la structure creuse des matrices L_r, L_q et M montrer que la matrice $\mathcal{A}$ a la structure suivante :

$$\mathcal{A} = \begin{pmatrix} D_0 & B_0 & 0 & \cdots & C_0 \\ C_1 & D_1 & B_1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & C_{N-2} & D_{N-2} & B_{N-2} \\ B_{N-1} & \cdots & 0 & C_{N-1} & D_{N-1} \end{pmatrix} \tag{20}$$

Q-4-3 : En déduire que trois vecteurs A, B, C de taille N suffisent pour le stockage de $\mathcal{A}$.

On s'évertue à réduire au maximum le coût en temps de calcul.

On va se servir d'un outil très utilisé en **optimisation numérique**, dénommé formule de **Sherman-Morrison**, qui permet de déterminer l'inverse d'une perturbation de rang 1 d'une matrice inversible.

Q-4-4 : Soit B une matrice carrée inversible de taille n et u, v deux vecteurs colonnes de taille n. Montrer qu'on a le résultat suivant :

$$(B + uv^T)^{-1} = B^{-1} - \frac{B^{-1}uv^TB^{-1}}{1 + v^TB^{-1}u} \tag{21}$$

Q-4-5 : Justifier le bien fondé de l'écriture équivalente suivante de la formule (21) :

$$(B + uv^T)^{-1} = \big(I_n - \frac{(B^{-1}u)v^T}{1 + v^T(B^{-1}u)}\big)B^{-1} \tag{22}$$

où I_n est la matrice identité de taille n.

On s'attaque maintenant à une résolution efficace d'un système linéaire avec la matrice $\mathcal{A}$. Ceci est plus judicieux que son inversion à proprement dit.

Soit Γ un réel tel que $D_0 - \Gamma$ et $D_{N-1} - B_{N-1}\Gamma^{-1}C_0$ sont non nuls.

On définit les vecteurs

$$U^T = \begin{pmatrix} \Gamma^T & 0 & \cdots & 0 & B_{N-1}^T \end{pmatrix}$$
$$V^T = \begin{pmatrix} 1 & 0 & \cdots & 0 & \Gamma^{-1}C_0 \end{pmatrix},$$

Q-4-6 : Montrer que $\mathcal{A} = \mathcal{B} + UV^T$ avec

$$\mathcal{B} = \begin{pmatrix} D_0 - \Gamma & B_0 & 0 & \cdots & 0 \\ C_1 & D_1 & B_1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & C_{N-2} & D_{N-2} & B_{N-2} \\ 0 & \cdots & 0 & C_{N-1} & D_{N-1} - B_{N-1}\Gamma^{-1}C_0 \end{pmatrix},$$

Q-4-7 : La matrice $\mathcal{B}$ étant maintenant une matrice tridiagonale, en proposant une factorisation LU avec stockage optimal de $\mathcal{B}$, proposer un algorithme de résolution de système linéaire avec la matrice $\mathcal{A}$.

Q-4-8 : En déduire une évaluation efficace du second membre $F(t, U)$ du système d'EDO.

Partie - 5 *Plan de validation numérique*

On définit convenablement un protocole de validations numériques qu'on réalise ensuite

A ce stade il est question de définir les outils de mesure qui serviront à quantifier efficacement l'ensemble des travaux réalisés dans les précédentes sections. Il est question non seulement de se convaincre soi-même mais aussi de convaincre le lecteur. Cela nécessite un choix optimal des tests à réaliser. Ici encore un travail de modélisation peut voir jour ainsi qu'un effort non négligeable de mise en oeuvre informatique.

Pour quantifier la qualité de notre développement, nous allons comparer la solution calculée à une solution exacte. Une solution exacte pour le problème considéré (1) est ce que nous avons désigné comme **onde solitaire**, dont l'expression est donnée par la formule (5), dans laquelle on désigne par **amplitude** la quantité $A = 3(C_s - 1)$.

On définit la fonction suivante, qui permet de mesurer l'erreur commise :

$$\mathcal{E}(t) = \frac{\int_a^b \left(u(t,x) - u_h(t_n,x)\right)^2 \, dx}{\int_a^b u^2(0,x)\, dx}, \tag{23}$$

où $u(t,x)$ est la solution exacte, et $u_h(t_n,x)$ est la solution calculée à l'instant t_n.

Cette définition peut prêter à confusion, mais les ambiguïtés seront levées dans ce qui suit. En effet, l'erreur définie par la mesure de (23) à l'instant t_n, ne saurait être suffisante pour quantifier la qualité de l'approximation numérique d'une onde solitaire. Pour cela on introduit d'autre moyens de mesure comme expliqué ci-après : On commence par introduire un instant t^* proche de t_n où $\mathcal{E}(t)$ atteint son extrémum (la détermination de cet instant t^* est un autre problème à résoudre comme nous l'évoquions). Puis on introduit les mesures suivantes :

- **Erreur de forme à l'instant** t_n : désignée et définie par $\mathbf{E^n} = \mathcal{E}(\mathbf{t}^*)$.
- **Erreur de phase à l'instant** t_n : désignée et définie par $\mathbf{P^n} = \mathbf{t_n} - \mathbf{t}^*$.
- **Erreur d'amplitude à l'instant** t_n : désignée et définie par $\mathbf{A^n} = (\mathbf{A} - \max_{\mathbf{a \le x \le b}} \mathbf{u_h(t_n, x)})/\mathbf{A}$.

Q-5-1 : En utilisant une méthode de Newton, déterminer t^* puis évaluer les différentes erreurs définies ci-dessus. Ecrire le résultat de manière formatée dans un fichier texte. Un exemple de formatage pourra consister à faire apparaître sur une colonne les instants t_n et sur les colonnes suivantes les différentes erreurs définies ci-dessus, en notations scientifiques avec suffisamment de chiffres significatifs.

Q-5-2 : Représenter graphiquement les valeurs exactes et calculées des quantités $I_1(t)$ et $I_2(t)$ en fonction de t.

Q-5-3 : Commenter les résultats obtenues.

Partie - 6 *Application réelle*

On s'adresse enfin à un problème concret

A ce stade on est convaincu de la qualité des développements effectués. On peut alors s'attaquer à des applications concrètes.

Q-6-1 **:** On considère ici la **collision de deux ondes solitaires**.

En prenant les données suivantes :

$f(u) = \frac{u^2}{2}$, $a = -20$, $b = 20$, $t_0 = 0$, $T = 10$, $\mu = \varepsilon = 5 \cdot 10^{-4}$
$u_0(x) = u(0, x + 0.5, 1.3, 0.5) + u(0, x - 0.5, 1.1, 0.5)$, (u est donnée par (5)).

Fournir trois graphiques, illustrant la solution du problème respectivement avant, pendant et après la collision des deux ondes solitaires. Commentez les résultats obtenus.

Q-6-2 **:** On considère ici les **limites dispersives**.

On souhaite observer la solution de l'équation (1) lorsque les paramètres μ et ϵ tendent vers zéro. C'est une configuration qui nécessite des pas de maillage h assez petits afin d'être en mesure de capturer les phénomènes physiques attendus. C'est ici que l'intérêt du stockage optimal de la matrice $\mathcal{A}$ et son inversion efficace voient leur importance.

— On considère le cas où $\varepsilon \neq 0, \mu = 0$: représenter la solution numérique de (1)

$$\mu = 0, \;\; \varepsilon = 10^{-6}, \;\; f(u) = \frac{u^2}{2}, \;\; a = 0, \;\; b = 1, \;\; t_0 = 0, \; T = 0.5, \tag{24}$$

et

$$u_0(x) = 2 + 0.5 \sin(2\pi x). \tag{25}$$

— On considère maintenant un cas où $\varepsilon = 0, \mu \neq 0$: représenter la solution numérique de (1)

$$\mu = 10^{-6}, \;\; \varepsilon = 0, \;\; f(u) = \frac{u^2}{2}, \;\; a = 0, \;\; b = 1, \;\; t_0 = 0, \; T = 0.5, \tag{26}$$

et

$$u_0(x) = 2 + 0.5 \sin(2\pi x). \tag{27}$$

— Commenter les résultats obtenus.

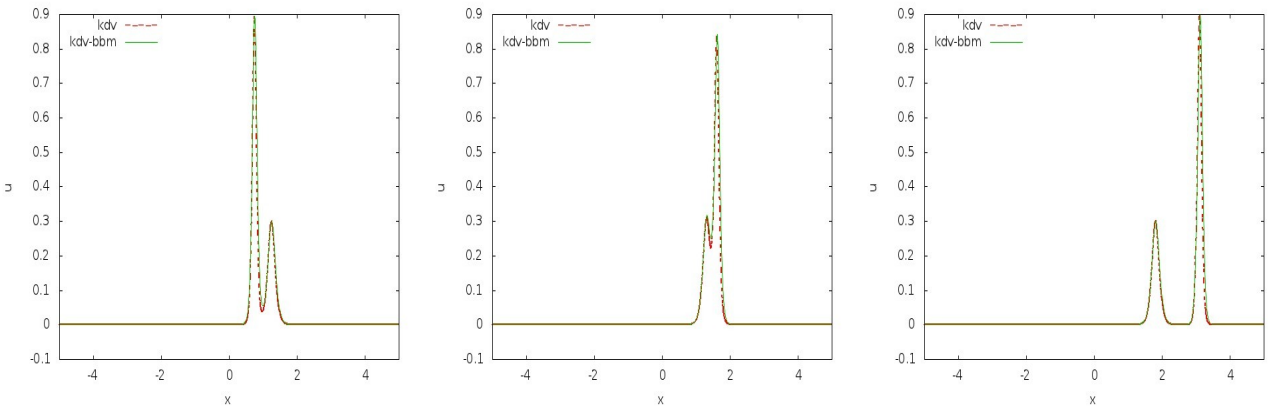

FIGURE 1 – Collision de deux ondes solitaires : avant la collision (à gauche), pendant la collision (au centre), après la collision (au centre).

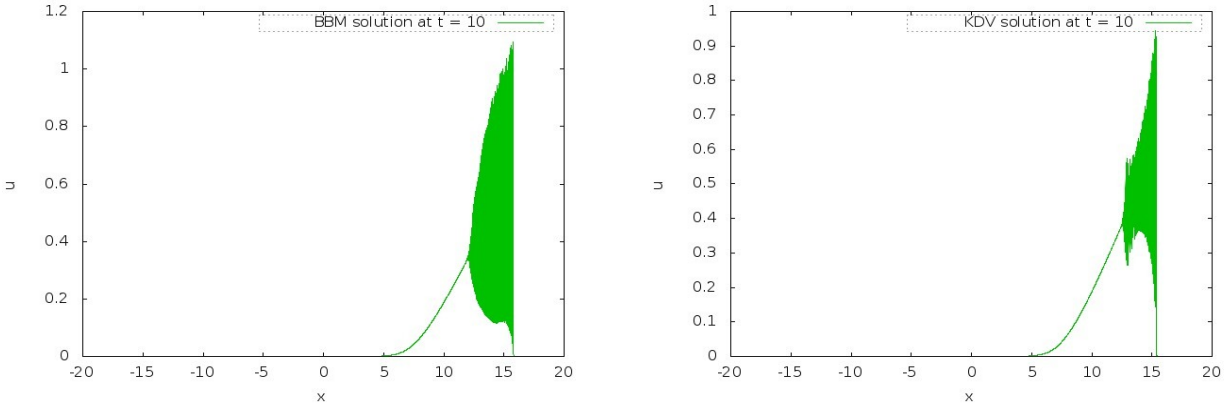

FIGURE 2 – Limites dispersives à un instant $t = 10s$ pour les paramètres fournis en exercice : modèle BBM (à gauche) et modèle KDV (à droite).

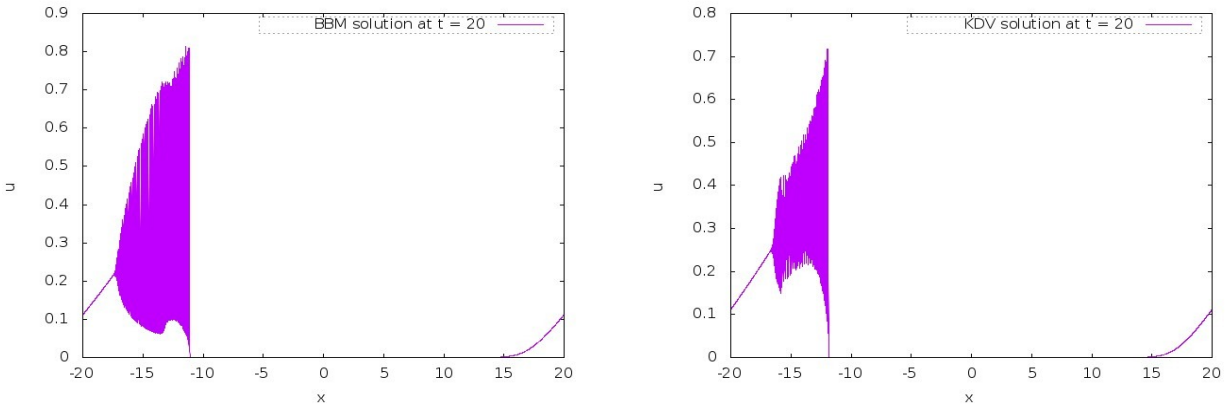

FIGURE 3 – Limites dispersives à un instant $t = 20s$ pour les paramètres fournis en exercice : modèle BBM (à gauche) et modèle KDV (à droite).